SAM Design

Adam Lyon

Fermilab Computing Division

August 3, 2006

Outline

1. Process and Job ID

2. Testing & Releases

3. Backwards compability of SQL Builder

4. Retired Files

Process and Job ID

Purpose

Provide a link between the consumer process ID and the batch job ID for easy file recovery.

- Output files lost due to worker node failure, network errors, pilot error.
- No easy way to recover these files.
- While the system does know what files were consumed by a consumer process (`CONSUMED_FILES.PROCESS_ID`), the user typically does not know the CPID associated with the lost output file.
- But the user does know the project name and a batch job ID associated with the lost output file.
- If there was a connection between the batch job ID and the CPID, a recovery could be made.

Implementing a Process to Batch Job Connection

- Need to put the batch job ID into the database to connect to the CPID.
- Note that the batch job ID must be a string (to accomodate DØ).
- For CDF, the batch job ID is actually a job ID and a section ID. But, project name and section ID will identifiy a section.
- Where to put the batch job information?
- `BATCH_JOBS_*` tables?
  - Seem to be for the old DØ Farm Request system.
  - Only connection to Consumer ID, not Consumer Process ID
  - Batch job ID is numeric.
- `PROCESSES` tables?
  - Since one to one mapping between process and batch job, is an obvious place.
  - Add a column? Then new schema release and IDLs have to change.
  - Use the `PROCESS_DESC` column?
  - Use the `LOCAL_PROCESS_ID` column? Seems to be the unix pid of the process. Is a numeric column.

Using PROCESSES.PROCESS_DESC

Seems to be filled in with a short description of the job (automatically by the experiment framework?)

CDF - 2,213,336 rows

Text	Count
demo	8767
C++ API Test Process	3
SAM Consumer Process	505,298
Consumer Process	45
Blank	1,699,223

Is anyone looking at this information?
Probably not!

DØ- 10,586,995 rows

Text	Count
framework process	586,400
none	12
test process	13
C++ API Test Process	136
test reco	7
Consumer Process	699,843
Offline Calibration DB Test	44
SAMcppClient Process	12,790
demo process	12
L3 primary vertexing tool	17
analysis	185
D0 Framework Process	125,264
Blank	9,162,272

Using PROCESSES.PROCESS_DESC II

- Good
  - Current information is mostly useless.
  - It is a string field!
  - Modification and access functions must already exist (maybe even in the C++ API). IDL must already exist.
  - Information is already printed in the project dump.
- Bad
  - We are overloading this field.
  - We have to find out who is filling it currently (`sam_manager`, `disk_cache_i`).
- TODO
  - Figure out where to fill information.
  - How to handle difference in CDF/DØ job names?
  - Do CDF people know the job and section ID, or just the section ID?
  - Write a dimension to access.

Testing & Releases

- We need a slighlty more formal release procedure in order to...
  - Make testing easier and more automatic.
  - Put release notes in a common place (so can be easily evaluated by experiments).
  - Make releases in a timely fashion (but not overly timely).
- Do we want to schedule releases?
  - I think this is somewhat awkward - leads to rushed development or development waiting for a long time.
  - But makes releases predictible.
- We should be adding tests for new features added or for bugs fixed.

Backwards compability of the SQL Builder

- As we know, the current dimensions parser is a big mess.
- Keeping old queries working exactly as they did before in the new system means...
  - Replicating features (and bugs) of the old system in the new system.
  - Requires a deep understanding of the old system - perhaps is impossible.
- Options?
  - Code the new system as it **should** work (only replicating the obvious and big features of the old system).
  - Old queries that now behave differently are ignored (if someone took advantage of a bug, it's their fault).
  - Somehow run the old system for old queries, run the new system for new queries (so old queries don't change).

Retired Files

- RETIRED_DATE column has been added to DATA_FILES. It is now only in CDF and DØ development databases.
- What do we need to do to put it into a production?
  - Come up with a retirement procedure (straight SQL for now - IDL and sam db server/client commands later?).
  - Test this procedure in development (can add a file with the same name but no, or different retirement date).
  - Are there unintended consequences or problems (is file name uniqueness checked somewhere else)?